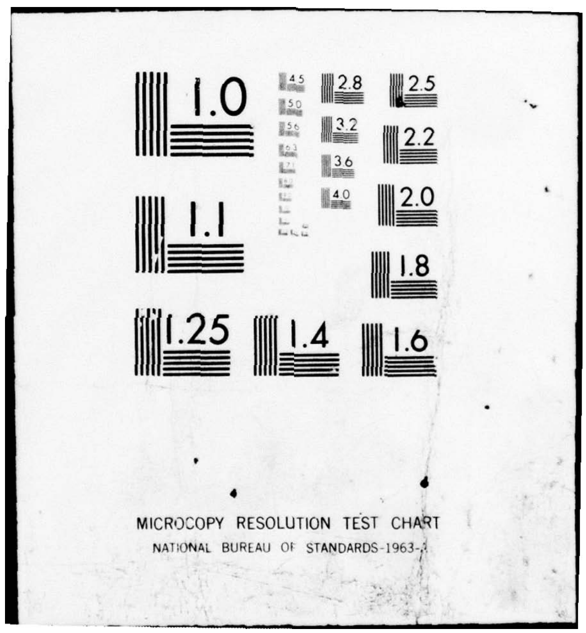
AD-A071 201 SOUTHERN METHODIST UNIV DALLAS TEX DEPT OF STATISTICS F/G 12/1
A RELATIONSHIP BETWEEN GENERALIZED AND INTEGRATED MEAN SQUARED --ETC(U)
1979 R F GUNST, J L HESS AFOSR-75-2871

UNCLASSIFIED AFOSR-TR-79-0811 NL

1 OF 1

AD
A071201

END
DATE
FILMED
8 --79
DDC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER AFOSR-TR- 79-0811	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) A RELATIONSHIP BETWEEN GENERALIZED AND INTEGRATED MEAN SQUARED ERRORS		5. TYPE OF REPORT & PERIOD COVERED Interim rept.
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) R.F. Gunst and J.L. Hess		8. CONTRACT OR GRANT NUMBER(s) AFOSR-75-2871
9. PERFORMING ORGANIZATION NAME AND ADDRESS Southern Methodist University Department of Statistics Dallas, Texas 75275		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 61102F 2304/A5
11. CONTROLLING OFFICE NAME AND ADDRESS Air Force Office of Scientific Research/NM Bolling AFB, Washington, DC 20332		12. REPORT DATE 1979
		13. NUMBER OF PAGES 6
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

SUBMITTED TO COMMUNICATION IN STATISTICS

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Regression Models
Biased Estimation
Generalized Mean Squared Error
Integrated Mean Squared Error

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Generalized mean squared error is a flexible measure of the adequacy of a regression estimator. It allows specific characteristics of the regression model and its intended use to be incorporated in the measure itself. Similarly, integrated mean squared error enables a researchr to stipulate particular regions of interest and weighting functions in the assessment of a prediction equation. The appeal of both measures is their ability to

DD FORM 1473

404 198

AFOSR-TR- 79-0811

A RELATIONSHIP BETWEEN GENERALIZED AND INTEGRATED MEAN SQUARED ERRORS

J. L. Hess
Kansas State University
Manhattan, Kansas U.S.A.

and

R. F. Gunst
Southern Methodist University
Dallas, Texas U.S.A.

SUMMARY

Generalized mean squared error is a flexible measure of the adequacy of a regression estimator. It allows specific characteristics of the regression model and its intended use to be incorporated in the measure itself. Similarly, integrated mean squared error enables a researcher to stipulate particular regions of interest and weighting functions in the assessment of a prediction equation. The appeal of both measures is their ability to allow design or model characteristics to directly influence the evaluation of fitted regression models. In this note an equivalence of the two measures is established for correctly specified models.

Keywords: Regression models; biased estimation; generalized mean squared error; integrated mean squared error

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH (AFSC)
NOTICE OF TRANSMITTAL TO DDC
This technical report has been reviewed and is
approved for public release IAW AFR 190-12 (7b).
Distribution is unlimited.
A. D. BLOSE
Technical Information Officer

1. INTRODUCTION

Define a multiple linear regression model as

$$\underline{Y} = \beta_0 \underline{1} + X\underline{\beta} + \underline{\varepsilon} \qquad (1)$$

where $\underline{Y}$ is an $(n\times 1)$ vector of response variables, $\underline{1}$ is an $(n\times 1)$ vector of ones, X is an $(n\times p)$ full-column-rank matrix of known constants (predictor variables) that is standardized so that $X'X$ is in correlation form, β_0 is an unknown constant, $\underline{\beta}$ is a $(p\times 1)$ vector of unknown regression coefficients, and $\underline{\varepsilon} \sim N(\underline{0},\sigma^2 I)$ is an $(n\times 1)$ vector of unobservable random error terms. We seek a prediction equation for the response variable of the form

$$\hat{Y}(\underline{u}) = \hat{\beta}_0 + \underline{u}'\hat{\underline{\beta}} \ , \qquad (2)$$

where $\underline{u}$ is a $(p\times 1)$ vector of standardized predictor variables, $\hat{\beta}_0 = \bar{Y}$, and $\hat{\underline{\beta}}$ is a suitable estimator of $\underline{\beta}$. By far the most often utilized estimator of $\underline{\beta}$ is the least squares estimator.

Biased regression estimators are popular alternatives to least squares when predictor variables are multicollinear. Two widely-used biased estimators are the principal component (Massy (1965), Marquardt (1970)) and (simple) ridge regression (Hoerl and Kennard (1970)) estimators. These estimators are generally compared with least squares and one another on the basis of their (total) mean squared errors,

$$T = E[(\hat{\underline{\beta}} - \underline{\beta})'(\hat{\underline{\beta}} - \underline{\beta})]. \qquad (3)$$

Responding to criticism that this criterion is an overly restrictive measure of the adequacy of a regression estimator, Theobald (1974)

investigated generalized mean squared errors of least squares and ridge regression estimators. Generalized mean squared error is defined as

$$G = E[(\hat{\underline{\beta}} - \underline{\beta})'M(\hat{\underline{\beta}} - \underline{\beta})], \quad (4)$$

for nonnegative definite matrices M. The addition of the matrix M in the definition of generalized mean squared error allows a researcher the flexibility of stressing specific linear combinations or individual elements of $\underline{\beta}$ more heavily than others in the computation of G.

Theobald (1974) showed that a range of values of k, the ridge parameter, exists which guarantees that the ridge estimator has smaller generalized mean squared error than least squares for all choices of M. He also found sufficient conditions for the ridge estimator to have smaller generalized mean squared error than least squares for all M. The implication of this result is that if the ridge parameter satisfies the sufficient conditions, all linear combinations of the regression coefficients can be estimated with smaller mean squared error using ridge regression than least squares.

Extending Theobald's results, Gunst and Mason (1976) derived sufficient conditions for the principal component estimator to have smaller generalized mean squared error than least squares for all M. There is not an existence theorem similar to that for ridge regression which guarantees that a principal component estimator can always be found which has smaller generalized mean squared error than least squares for all choices of M. Gunst and Mason (1976) also show that neither principal components nor ridge regression dominates the other in generalized mean squared error for all choices of M.

Box and Draper (1959) stimulated a series of papers by several authors on the use of integrated mean squared error as a criterion for evaluating various designs for fitting response surface models. Integrated mean squared error is defined as

$$J = \int_R \cdots \int E\{(\hat{Y}(\underline{u}) - E[Y(\underline{u})])^2\} \, W(\underline{u}) d\underline{u}. \tag{5}$$

As defined in (5), integrated mean squared error incorporates the mean squared error of the prediction equation at a point $\underline{u} \in R$, weighted by a function $W(\underline{u})$, and integrated over a region R of interest to the researcher. Since the researcher can define a weight function and region of interest to suit the purpose of his investigations, integrated mean squared error offers the type of generality and flexibility in the evaluation of prediction equations as generalized mean squared error does for regression estimators.

Gunst and Mason (1979) adopted integrated mean squared error as a criterion for the assessment of prediction equations for which the matrix of predictor variables is fixed but one can choose which regression estimator to use. Specifically, integrated mean squared error was employed to evaluate competing prediction equations based on least squares, principal component, and ridge regression estimators when the matrix of predictor variables is multicollinear. We now wish to identify conditions for which the two model-fitting criteria are equivalent.

2. AN EQUIVALENCE RELATIONSHIP

Helms (1971) shows that integrated mean squared error, J, is affected by the weight function and the region of interest only through the second order moment matrix

$$\ddagger = \int_R \cdots \int \underline{u}\underline{u}' W(\underline{u}) d\underline{u} \, . \qquad (6)$$

We now propose the following theorem.

Theorem: If a multiple linear regression model is correctly specified as in (1),

$$J = n^{-1}\sigma^2 + G$$

when $M = \ddagger$.

Proof:
$$J = \int_R \cdots \int E\{(\hat{Y}(\underline{u}) - E[Y(\underline{u})])^2\} \, W(\underline{u}) d\underline{u}$$
$$= E\{\int_R \cdots \int (\bar{Y} - \beta_0)^2 \, W(\underline{u}) \, d\underline{u}\}$$
$$+ 2E\{\int_R \cdots \int (\bar{Y} - \beta_0)(\underline{u}'\hat{\underline{\beta}} - \underline{u}'\underline{\beta}) \, W(\underline{u}) d\underline{u}\}$$
$$+ E\{\int_R \cdots \int (\underline{u}'\hat{\underline{\beta}} - \underline{u}'\underline{\beta})^2 W(\underline{u}) d\underline{u}\}$$
$$= n^{-1}\sigma^2 + E\{(\hat{\underline{\beta}} - \underline{\beta})'\ddagger(\hat{\underline{\beta}} - \underline{\beta})\}.$$

An important consequence of this theorem is that the existence and sufficient conditions established by Theobald (1974) for the ridge estimator to have smaller generalized mean squared error than least squares for all choices of M now guarantee that under those same conditions the integrated mean squared error of the ridge prediction equation is less than that for least squares with all choices of $W(\underline{u})$ and R that produce nonnegative definite second order moment matrices. Similar equivalence relationships hold for the generalized mean squared error properties established in Gunst and Mason (1976).

ACKNOWLEDGEMENT

This research was supported in part by the Air Force Office of Scientific Research, Air Force Systems Command, USAF, under grant No. AFOSR-75-2871.

REFERENCES

Box, G.E.P. and Draper, N.R. (1959). A basis for the selection of a response surface design, J. Amer. Statist. Assn., 54, 622-54.

Gunst, R.F. and Mason, R.L. (1976). Generalized mean squared error properties of regression estimators, Commun. In Statist., A5, 1501-8.

Gunst, R.F. and Mason, R.L. (1979). Some considerations in the evaluation of alternate prediction equations, Technometrics, 21, 55-63.

Helms, R.W. (1971). The predictors average estimated variance criterion for the selection-of-variables problem in general linear models, Department of Biostatistics, University of North Carolina at Chapel Hill, Institute of Statistics Mimeo Series #777.

Hoert, A.E. and Kennard, R.W. (1970). Ridge regression: biased estimation for non-orthogonal problems, Technometrics, 12, 55-67.

Marquardt, D.W. (1970). Generalized inverses, ridge regression, biased linear estimation and nonlinear estimation, Technometrics, 12, 591-612.

Massy, W.F. (1965). Principal component regression in exploratory statistical research, J. Amer. Statist. Assn., 60, 234-56.

Theobald, C.M. (1974). Generalizations of mean square error applied to ridge regression, J.R. Statist. Soc. B, 36, 103-6.